specifying interactions of ILK with downstream, cytoplasmic or cytoskeletal proteins. Reduced ECM adhesion by the p59^{ILK} overexpressing cells is consistent with our observation of adhesion-dependent inhibition of ILK activity, and suggests that p59^{ILK} plays a role in inside-out integrin signalling. Furthermore the p59^{ILK}-induced, anchorage-independent growth of epithelial cells indicates a role for ILK in mediating intracellular signal transduction by integrins[14-16]. □

Received 15 August; accepted 30 October 1995.

1. Damsky, C. H. & Werb, Z. *Curr. Opin. Cell Biol.* **4,** 772–781 (1992).
2. Hynes, R. O. *Cell* **69,** 11–25 (1992).
3. Clark, E. A. & Brugge, J. S. *Science* **268,** 233–239 (1995).
4. Fields, S. & Song, O. *Nature* **340,** 245–246 (1989).
5. Lux, S. E., John, K. M. & Bennett, V. *Nature* **344,** 36–42 (1990).
6. Inoue, J.-I. *et al. Proc. natn. Acad. Sci. U.S.A.* **89,** 4333–4337 (1992).
7. Lukas, J. *et al. Nature* **375,** 503–506 (1993).
8. Schaller, M. D. *et al. Proc. natn. Acad. Sci. U.S.A.* **89,** 5192–5196 (1992).
9. Hanks, S. K., Calalb, M. B., Harper, M. C. & Patel, S. K. *Proc. natn. Acad. Sci. U.S.A.* **89,** 8481–8491 (1992).
10. Dedhar, S., Saulnier, R., Nagle, R. & Overall, C. M. *Clin. exp. Metastasis* **11,** 391–400 (1993).
11. Chen, Y.-P. *et al. J. biol. Chem.* **269,** 18307–18310 (1994).
12. Filmus, J. *et al. Oncogene* **9,** 3627–3633 (1994).
13. O'Toole, T. E. *et al. J. Cell Biol.* **124,** 1047–1059 (1994).
14. Kapron-Bras, C., Fitz-Gibbon, L., Jeevaratnam, P., Wilkins, J. & Dedhar, S. *J. biol. Chem.* **268,** 20701–20704 (1993).
15. Chen, Q., Kinch, M. S., Lin, T. H., Burridge, K. & Juliano, R. L. *J. biol. Chem.* **269,** 26602–26605 (1994).
16. Schlaepfer, D. D., Hanks, S. K., Hunter, T. & van der Geer, P. *Nature* **372,** 786–791 (1994).
17. Kozak, M. *Cell* **44,** 283–292 (1986).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. molec. Biol.* **215,** 403–410 (1990).
19. Hanks, S. K., Quinn, A. M. & Hunter T. *Science* **241,** 42–52 (1988).
20. Zervos, A. S., Gyuris, J. & Brent, R. *Cell* **72,** 223–232 (1993).
21. Argraves, W. S. *et al. J. Cell Biol.* **105,** 1183–1190 (1987).
22. Gietz, D., St Jean, A., Woods, R. A. & Schiestl, R. H. *Nucleic Acids Res.* **20,** 1425 (1992).
23. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory Press, New York, 1989).
24. Otey, C. A., Pavalko, F. M. & Burridge, K. *J. Cell Biol.* **111,** 721–729 (1990).
25. Cooper, J. A., Sefton, B. M. & Hunter, T. *Meth. Enzym.* **99,** 387–402 (1983).
26. Stephens, L. C., Sonne, J. E., Fitzgerald, M. L. & Damsky, C. H. *J. Cell Biol.* **123,** 1607–1620 (1993).
27. Harlow, E. & Lane, D. *Antibodies: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, New York, 1988).
28. Leung-Hagesteijn, C. Y., Milankov, K., Michalak, M., Wilkins, J. & Dedhar, S. *J. Cell Sci.* **107,** 589–600 (1994).
29. Brick, R. N., Filmus, J. & Quaroni, A. *Expl Cell Res.* **170,** 300–309 (1987).

ACKNOWLEDGEMENTS. We are grateful to R. Brent and R. Finley (Massachusetts General Hospital) for providing the yeast strains and two-hybrid plasmids, and to A. Murthy (MGH) for discussions regarding yeast transformations. Thanks to C. Otey (Univ. of Virginia) for providing the synthetic integrin cytoplasmic domain peptides, Y. Ben-David and S.-J. Lu (Sunnybrook) for the WM35 melanoma cDNA library, and M. Park (McGill University) for the placental cDNA library. Monoclonal antibodies to β_1 integrin were provided by C. Damsky (Univ. of California at San Francisco) and J. Wilkins (Univ. of Manitoba). We appreciate critical reading of the manuscript by J. Woodgett (Univ. of Toronto). G.H. was the recipient of postdoctoral fellowships from the Medical Research Council of Canada and the Sunnybrook Trust for Medical Research. S.D. is a Terry Fox Scientist of the National Cancer Institute of Canada, and is supported by grants from the NCIC and MRC(C).

Conserved residues and the mechanism of protein folding

E. Shakhnovich, V. Abkevich & O. Ptitsyn*

Harvard University, Department of Chemistry, 12 Oxford Street, Cambridge, Massachusetts 02138, USA
* Institute of Protein Research, Russian Academy of Sciences, Puschino, Moscow Region 142292, Russia, and NIH, National Cancer Institute, Molecular Structure Section, LMMB, Bethesda, Maryland 29892-3677, USA

EXPERIMENTAL[1-4] and simulation[7] studies show that small monomeric proteins fold in one kinetic step, which entails overcoming the free-energy barrier between the unfolded and the native protein through a transition state[8,9]. Two models of transition state formation have been proposed: a 'nonspecific' one in which it depends on the formation of a sufficient number of native-like contacts regardless of what amino acids are involved[10-12], and a 'specific' one, in which it depends on formation of a specific subset of the native structure (a folding nucleus)[8,13,14]. The latter requires that some amino acids form most of their contacts in the transition state, whereas others only do so on reaching the native conformation. If so, mutations affecting the stability of the transition state nucleus should have a greater effect on the folding kinetics than mutations elsewhere, and the residues involved should be evolutionarily conserved. Lattice-model simulations and experiments[8,13-16] suggest that such mutations exist. Here we present a method for determining the folding nucleus of a protein with known structure with two-state folding kinetics. This method is based on the alignment of many sequences designed to fold into the native conformation of a protein to identify the positions where amino acids are most conserved in designed sequences. The method is applied to chymotrypsin inhibitor 2 (CI2), a protein whose transition state has been previously studied by protein engineering[14-16]. The involvement of residues in folding nucleus of CI2 is clearly correlated with their conservation in design, and the residues forming the nucleus are highly conserved in 23 natural sequences homologous to CI2.

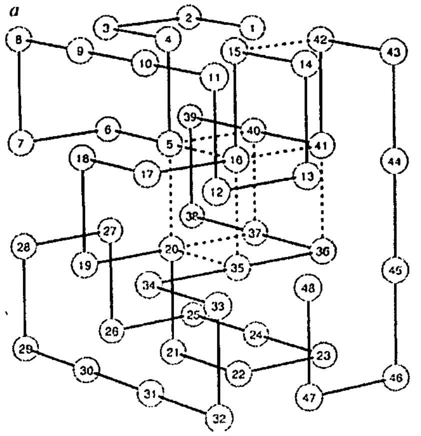
We first studied a simple lattice model of the protein[17]. (1) We chose an arbitrary conformation of lattice protein chain to serve as the native structure and then (2) selected sequences that deliver low energy in this 'native' structure compared to unfolded and misfolded conformations. (3) The designed sequences (from 2) were folded using Monte Carlo simulations. Because folding simulations and sequence design are carried out using the same set of potentials, a self-consistent study of the model can be carried out with any choice of potential. For some (but not all) predictions on real proteins it is possible that the particular choice of potentials is not very important[17] (see below).

Following this method we chose the target conformation (Fig. 1), and designed sequences to fold to it. Further, the analysis based on Monte Carlo folding simulations[6] revealed the folding nucleus (Fig. 1*a*) for both sequences shown in Fig. 1*b*,*c*.

The stochastic design algorithm generated 10^6 sequences, each having low energy in the conformation shown in Fig. 1*a*. Alignment of these sequences revealed a remarkable feature: correlation between residue conservatism and its participation in the folding nucleus (Fig. 2). Indeed, all four most conserved residues (5,16,20,35) belong to the nucleus (see Fig. 1*a*). This could have been due to the degree they are buried in the native structure (see Fig. 1), but this is not the case because the conservation of all buried residues is smaller than that of nucleus residues (see Fig. 2*b*).

The key feature of the nucleation mechanism is the existence of 'kinetically important' positions. To test this we designed a set of sequences where positions (5,16,20,35) inside the nucleus are constrained to 'alanines', which destabilize the nucleus (in an MJ parameter set). We studied the folding kinetics for these sequences (Fig. 3). The existence of some 'alanine' residues in the nucleus caused a pronounced slowing down of folding. This is an exclusively kinetic effect: several 'wild-type' sequences (designed without forcing 'alanines' into the nucleus) have native-state energy comparable to that of sequences designed with alanines in the nucleus. This example implies that the present method of identifying a folding nucleus would give wrong predictions for sequences with alanines designed in the nucleus, because these sequences do not appear to fold through a specific nucleus mechanism. Therefore the present method only applies to fast-folding sequences.

We applied this method to predict the folding nucleus of CI2, originally studied in refs 15, 16. To this end we used the real off-lattice structure of CI2 as the target and designed sequences that have low energy in this conformation. There is a clear qualitative correlation between site conservation and ϕ-values for folding. Our calculations (see Fig. 4) predict that most conserved residues, particularly A35, I39, L68, I70 and I76, are likely to belong to the folding nucleus. After this work was completed, we learned from A. Fersht (personal communication) that A35 (not studied in refs



b
RQPDFYEQKDKTVERLRGGMGIATENYTNSASACILWHPFVDKLVSAL

c
PNCMDHASWGVGELFKGALKLSPTAIDILVQATFERDSNREQVKYYTR

FIG. 1 *a*, The conformation of the 48-mer chosen as the native state in our design/folding procedure. Each amino-acid residue is represented as a bead occupying a lattice site. Although the model does not treat side chains explicitly, the amino acids are chemically different; their differences are manifested in pairwise interaction energies of different magnitude and sign, depending on the identities of interacting amino acids. A conformation is described by the set of coordinates of all monomers $\{r\}$. The energy of a conformation is:

$$E = \sum_{i<j} B(\eta_i, \eta_j)\Delta(r_i - r_j) \qquad (1)$$

where $\Delta(r_i - r_j) = 1$ if monomers i and j are lattice neighbours not connected by a covalent bond, and 0 otherwise. $B(\eta, \xi)$ is the magnitude of interaction between amino acids of type η and ξ. There are 20 types of amino acid in the model. Two parameter sets, B, were used: one proposed by Miyazawa and Jemigan[19] (MJ), and the other proposed by Kolinski, Godzik and Skolnick[20] (KGS). The parameters given in Table 6 of ref. 19 were used as the MJ set. They represent the 'excess' pairwise interaction between two amino acids as compared with their averaged interaction with their environment. The values of B in equation (1) were shifted and normalized to achieve a zero average over all possible contacts and standard variance of unity. This amounts to multiplying all parameters in refs 19, 20 by a constant factor and adding a constant to each interaction parameter B. This procedure effectively sets the temperature scale for simulations (more details in ref. 8). For each set of parameters we designed sequences to fold to the conformation shown here. The design procedure has been described in detail elsewhere[17,21,22]. It is a stochastic (Monte Carlo) optimization routine in sequence space which keeps amino-acid composition unchanged. It minimizes energy of the native conformation. The condition of constant amino-acid composition makes it equivalent to optimizing the relative energy of the native state, or Z-score[23]. As is characteristic of Monte Carlo searches, unfavourable mutations can also be accepted, with a small probability, given by a Metropolis criterion[24] with selective temperature T_{sel}. We chose $T_{sel} = 0.15$ (in our temperature scale), which is sufficiently low to generate stable and fast-folding sequences. *b*, *c*, Two sequences designed to fold and be stable in the conformation shown in *a*: with KGS parameters (*b*), and with MJ parameters (*c*). The design tends to place the most strongly interacting amino acids in the interior where they can form most contacts. The strongest 'excess' interactions in MJ parameters from ref. 19 are between 'charged' groups (D and K) therefore they are buried in sequence (*c*). Alternatively, the strongest interactions in the KGS set are between hydrophobic groups, hence designs with these parameters yield more realistic sequences with hydrophobic groups buried, as in sequence *b*. Monte Carlo folding simulations were carried out for both sequences. Both of them folded to and were stable in the conformation shown in *a* with their respective forcefields. The nucleus (determined from the folding simulations, using the method described in detail in ref. 8) was identical for both sequences; it is shown by broken lines in *a*.

15, 16) is the residue most involved in the nucleus with $\phi \simeq 1.0$ (see also ref. 14): I76 looks like an exception with $\phi \simeq 0.13$. However, it is not: despite the low ϕ-value, recent measurements on a number of residues contacting I76 strongly suggest that I76 is also a key residue of the folding nucleus[14].

We also found a striking relationship between the residue conservatism in our design and evolutionary conservatism in the alignment of natural sequences homologous to CI2 (Fig. 4*b*)[18]. The key nucleus residue A35 is 100% conserved. Among the most conserved in Fig. 4*b* are nucleus residues I76, I39. Nucleus L68 is moderately conserved in the alignment, though, because of more frequent substitutions of L to V.

This finding also suggests that our design procedure might reflect on some features of the evolutionary process of protein morphogenesis, namely these aspects related to folding. It is possible that some of the surface residues are evolutionarily conserved for functional reasons not taken into account in the design procedure. To this end, the comparison of the results of the design (Fig. 4*a*) with the alignment (Fig. 4*b*) might be helpful in identifying which residues are conserved for 'folding' reasons and which ones are functionally conserved.

The possible physical rationale for the suggested method is that the sequence design identifies a contiguous cluster of core residues which are close enough to each other in space to form a contiguous network of interactions. The energy-based design

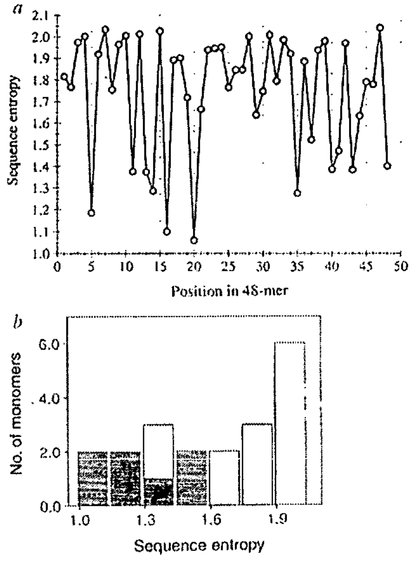


FIG. 2 *a*, The sequence design entropy is defined as:

$$S(i) = -\sum_{j=1}^{m} p_j(i)\ln p_j(i) \qquad (2)$$

where $p_j(i)$ is frequency of occurrence of amino acid of type j at site i; $m = 20$ is the total number of amino-acid types. To estimate the frequencies $p_i(j)$, we performed long runs of the Monte Carlo sequence design algorithm using a low selective temperature ($T_{sel} = 0.15$) to obtain $\sim 10^5$ sequences per parameter set. We estimated the desired frequencies as the fraction of this population of designed sequences bearing an amino acid of type j at position i. Shown is a plot of $S(i)$ as a function of position i for the KGS parameter set; the corresponding plot for the MJ parameter set is similar (not shown). *b*, Histogram for the distribution of design entropy for buried residues (having three or four non-covalent contacts in the native structure, Fig. 1*a*) from the plot in *a*. White bars, all buried residues; grey bars, those of buried residues that have two or more nucleus contacts.

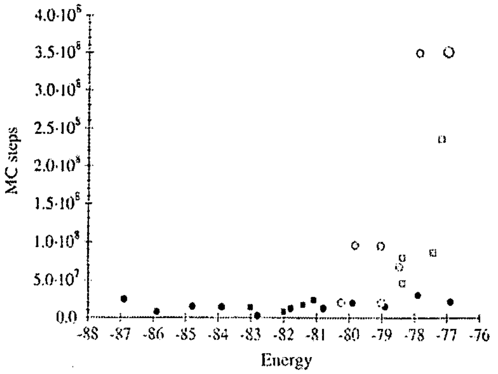


FIG. 3 The mean first-passage folding time for different sequences designed (with MJ parameters) to fold into the native structure shown in Fig. 1*a*. Monte Carlo folding simulations were done at a low temperature ($T = 0.8$) so that the folding time of the 'wild-type' sequences (black circles) would not depend dramatically on native-state energy[25]. This is indeed the case in the present simulations: 11 'wild-type' sequences having different energies in the native state (black circles) have similar average folding times. Low temperature for folding simulations was chosen to distinguish the special role of nucleus residues in folding from the more obvious[26-29] relation between folding rate and the stability of the native state. In the parameters from ref. 19, alanine interacts unfavourably with most other amino acids so that its placement in the nucleus destabilizes the latter. Several sequences were designed to have an alanine residue at the predetermined nucleus positions. (Alanine residues were placed at these positions and mutations were forbidden there, otherwise the design algorithm proceeded as usual.) Black squares correspond to sequences with one alenine residue, placed in either positions 5,16,20,35; grey circles correspond to sequences with 2 fixed alanine residues, in positions (5,16), (5,20), (16,20), (5,35), (16,35) or (20,35); grey squares correspond to sequences having 3 fixed alanines placed in all triplets out of positions (5,16,20.35), and the large grey circle corresponds to the sequence designed with fixed alanine residue at all four positions (5,16,20,35).

procedure conservatively places into these positions residues that strongly attract each other. This creates a relatively low free-energy set of partly folded conformations in which mutually stabilizing strong nucleus contacts are formed while other parts of the chain are disordered. In the model representing folding as kinetically two-state, such a set of conformations serves as a saddle-point in the free-energy landscape, which is the transition state. This also suggests that our current method may be applicable only to proteins with two-state folding kinetics.

Note added in proof: The notation for CI2 residues in ref. 14 is shifted by 19 units from the notation used in ref. 15 and this work; for example, A35 here would be A16 in ref. 14. □

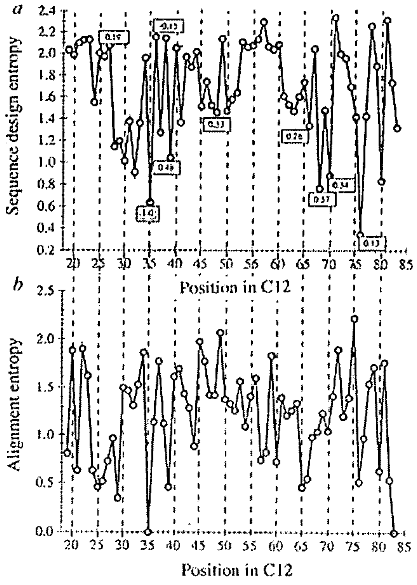


FIG. 4 The design entropy of CI2. The PDB-structure of CI2 (2ci2) was taken as the target conformation. Then, using an MC sequence design algorithm, we generated many sequences (with the same amino-acid composition as the native CI2) exhibiting a low energy in this target conformation. The energy function for design was calculated for C_β contact using equation (1). Two residues were considered to be in contact if the distance between their C_β atoms (C_α for G) was ≤ 7.5 Å and if they were more than two units apart from each other along the sequence. KGS and MJ parameters were used for $B(\eta_i, \eta_j)$ (see equation (1)). Amino-acid frequencies were taken from the designed sequences (as explained in Fig. 2) and substituted into equation (2). We compared these results with the numbers of native-like contacts of a given residue in the transition state relative to that in the native state (ϕ-values[30,31]). Labels represent the ϕ-values for all residues studied in refs 15, 16. The shaded label is the recent result of ref. 14 which only came to our attention after this study was completed. The plot shown here was calculated using the KGS parameters. The MJ parameters yielded very similar results indicating the same positions as conservative ones. Design with KGS parameters placed predominantly L and I in the conserved positions, which coincided in most cases (except position 35) with what is seen in real sequences. MJ parameters placed charged groups in these positions, for the reasons explained in the legend to Fig. 1. *b*, The alignment entropy (amino-acid variability) calculated over 23 sequences homologous to CI2 (ref. 18). Frequencies of amino-acid occurencies were taken from the alignment, and sequence alignment entropy was evaluated according to equation (2). (These data were already calculated in the file we used, 2ci2.hssp (ref. 18).)

Received 25 July; accepted 2 November 1995.

1. Jackson, S. E. & Fersht, A. R. *Biochemistry* **30,** 10428–10435 (1991).
2. Alexander, P., Orban, J. & Bryan, P. *Biochemistry* **31,** 7243–7248 (1992).
3. Schindler, T., Herrler, M., Marahiel, M. & Schmid, F. X. *Nature struct. Biol.* **2,** 663–673 (1995).
4. Sosnick, T. R., Mayne, L., Hiller, R. & Englander, S. W. *Nature struct. Biology* **1,** 149–156 (1994).
5. Viguera, A. R., Martinez, J. C., Filiminov, V. V., Mateo, P. L. & Serrano, L. *Biochemistry* **33,** 2142–2150 (1994).
6. Kragelund, B. B., Robinson, C. V., Knudson, J. & Dobson, C. M. *Biochemistry* **34,** 7117–7124 (1995).
7. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. *Biochemistry* **34,** 3066–3076 (1995).
8. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. *Biochemistry* **33,** 10026–10036 (1994).
9. Guo, Z. & Thirumalai, D. *Biopolymers* **35,** 137–139 (1995).
10. Shakhnovich, E. I., Farztdinov, G., Gutin, A. M. & Karplus, M. *Phys. Rev. Lett.* **67,** 1665–1668 (1991).
11. Sali, A., Shakhnovich, E. I. & Karplus, M. *Nature* **369,** 248–251 (1994).
12. Wolynes, P. G., Onuchic, J. N. & Thirumulai, D. *Science* **267,** 1619–1620 (1995).
13. Fersht, A. *Proc. natn. Acad. Sci. U.S.A.* **92,** 10869–10873 (1995).
14. Itzhaki, L., Otzen, D. & Fersht, A. *J. molec. Biol.* **254,** 260–288 (1995).
15. Jackson, S. E., elMasry, N. & Fersht, A. *Biochemistry* **32,** 11270–11278 (1993).
16. Otzen, D. E., Itzhaki, L., elMasry, N., Jackson, S. E. & Fersht, A. *Proc. natn. Acad. Sci. U.S.A.* **91,** 10422–10425 (1994).
17. Shakhnovich, E. I. *Phys. Rev. Lett.* **72,** 3907–3910 (1994).
18. Sander, C. & Schneider, R. *Proteins* **9** 56–68 (1991).
19. Myazawa, S. & Jernigan, R. *Macromolecules* **18,** 534–552 (1985).
20. Kolinski, A., Godzik, A. & Skolnick, J. *J. chem. Phys.* **98,** 7420–7433 (1993).
21. Shakhnovich, E. I. & Gutin, A. M. *Proc. natn. Acad. Sci. U.S.A.* **90,** 7195–7199 (1993).
22. Shakhnovich, E. I. & Gutin, A. M. *Prot. Engng* **6,** 793–800 (1993).
23. Bowie, J. U., Luthy, R. & Eisenberg, D. *Science* **253,** 164–169 (1991).
24. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. & Teller, E. *J. chem. Phys* **21,** 1087–1092 (1953).
25. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. *J. chem. Phys* **101,** 6052–6062 (1994).
26. Goldstein, R., Luthey-Schultzen, Z. A. & Wolynes, P. G. *Proc. natn. Acad. Sci. U.S.A.* **89,** 4918–4922 (1992).
27. Sali, A., Shakhnovich, E. I. & Karplus, M. *J. molec. Biol.* **235,** 1614–1636 (1994).
28. Hao, M.-H. & Scheraga, H. *J. chem. Phys* **102,** 1334–1339 (1995).
29. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. *Proc. natn. Acad. Sci. U.S.A.* **92,** 1282–1286 (1995).
30. Matouschek, A., Kellis, J., Serrano, L., Bycroft, M. & Fersht, A. *Nature* **346,** 440–445 (1990).
31. Matouschek, A., Serrano, L. & Fersht, A. *J. molec. Biol.* **224,** 819–835 (1992).

ACKNOWLEDGEMENTS. We are grateful to A. Fersht and M. Karplus for discussions and to G. Berriz for comments and editorial help. E.S. is supported by the Packard foundation. O.P. acknowledges support from the International Human Frontier Science Program. The alignment file 2ci2.hssp was obtained from the EMBL server, Heidelberg, Germany.